How Impactful Penalties Reflect the Language Directed Towards Referees

Indiana University

Scott Berry and Nick Erwin

Introduction

Referees in sports draw ire from fans on all levels of competition. While referees have messages directed at them throughout the entirety of the games, in particular, "big plays" can be heavily criticized and even be cause for concern for referees' safety after the games. Referees from little league to professional level regularly draw death threats, physical violence, and verbal abuse (Jacobs 2020). One recent instance of a referee getting attacked is from a small town called Edinburg, Texas where a high school football player brutally tackled a referee for a penalty call that he did not agree with. "It really is a stunning instance to see, and your heart goes out to the ref and his family" (ABC7, 2020). There have been several other instances of anger boiling over towards referees as well which makes it crucial to understand sentiment towards referees.

This study aims to determine how strongly big plays can affect language directed towards referees and if the sentiment could be at a dangerous level. This knowledge could help predict when a referee might need additional security after a game, as has proved to be necessary many times after competitions worldwide. The findings could support why referee bias leads to questionable decisions if they fear for their safety (Scoppa 2021). By looking deeper into the sentiment response of Reddit comments during the NFC Divisional game between the Tampa Bay Buccaneers and the New Orleans Saints, this study answers whether or not how big penalty calls during the game can affect the positive or negative response of the fans posting their opinions on Reddit. By correctly conducting this study, a new approach to how referees are treated during and after a game can be drastically improved from many of the instances we have seen in the past.

Research Question(s)

Normally, there are several penalties referees call during a game that either has positive reactions by one fanbase and usually a negative reaction from another. Referees are usually caught in the middle of these fanbases and will ultimately never be liked by everyone. Research has been conducted based on this topic to answer the following question.

- How strongly does a play's change in expected points probability indicate the sentiment of messages directed at referees?

Data

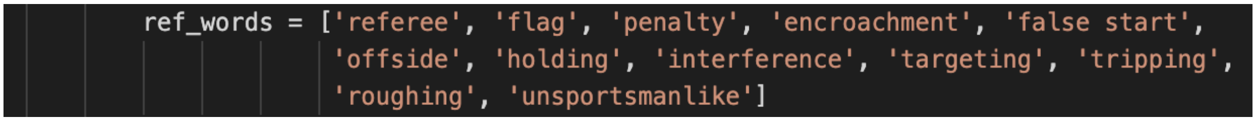
The data used for the study were the comments from the Reddit Game Thread of the NFC Divisional Game between the Tampa Bay Buccaneers and the New Orleans Saints (Reddit 2021). In the case of this study, any game could have been used for the analysis but decided to use the NFC Divisional Game because it was back and forth offensively and defensively for both teams the entire duration of the game. This allows for multiple fanbase inputs on the Reddit thread instead of a massively biased reaction from one side versus another. The NFC Divisional game also did not have any major controversial penalty calls from referees that went viral. This is a significant aspect of our data collection since a controversial call from a referee will lead to a higher number of comments about referees during a game. This game thread was selected due to being less than the PRAW maximum of 50,000 comments while still being a recent high-profile game. PRAW is the module to be utilized within a Python script to import all the Reddit comments and their metadata (Boe 2021). Once imported, the comment metadata is reformatted into a JSON structure with comments and their respective bodies, timestamps, and team flairs. The team flairs are included to associate comments to expected support for home and away teams. Timestamps from the comments are in the UNIX-timestamp format which will be used in the method later. The JSON structure is then saved as a file that will be referenced for data analysis. A total of 50,442 comments from the Reddit Game Thread were collected to be analyzed for our study.

To calculate penalty calls during the game, a Twitter account was used that posts play-by-play tweets during each game. Timestamps for the penalties were manually gathered from the @bucs_nation Twitter account by scrolling back through the account to the time of the game. Each penalty tweet had an associated time of day, such as 8:59 PM. 12 penalties occurred during the game and those penalties and timestamps were stored in a text file to be used in the method later. All penalties that occurred during the game will be imported using real-timestamps and then converted to a UNIX timestamp for consistency.

To determine a play's change in expected points probability, this study looked at the play-by-play information during the game provided by Pro Football Reference (Pro, 2021). This website offers detailed statistics about each game from 1920 to the present and is one of the most reliable sources online for accurate NFL information. This study looked at the 12 plays during the game where penalties were called and calculated the expected points delta from each play. Pro Football Reference shows expected points before and after the play and to calculate the delta from this information, the absolute value was used on the difference between points before and points after the play where a penalty was called.

Method

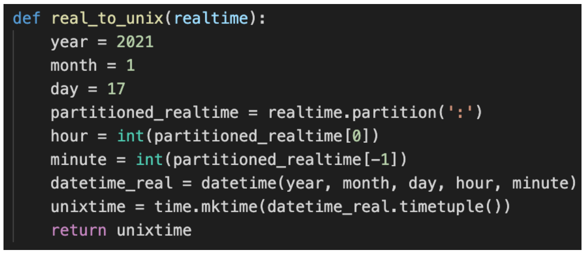
Since the goal of this study is to analyze sentiment towards referees for the penalties called during the game, the comments from the JSON file created from importing the Reddit Game Thread will be filtered down to only the comments that include words associated with referee criticism such as "referee", "penalty", and "flag". The purpose of these keywords is to remove the noise from other comments about the game that do not directly involve the referees. We then added a list of words for the most common penalties called during a football game. This includes words such as “holding”, “interference”, “offside”, etc (Rookie Road). This allows for filtering on the exact comments that reference plays where penalties were called and where referees were involved as well. With several keywords selected as shown in Figure 1, the final filter returned 1,628 relevant comments from the Reddit Game Thread.



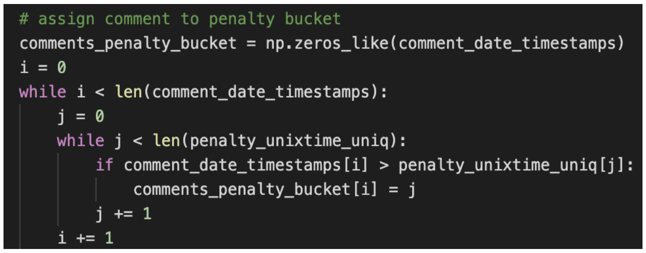
```
        ref_words = ['referee', 'flag', 'penalty', 'encroachment', 'false start',
                     'offside', 'holding', 'interference', 'targeting', 'tripping',
                     'roughing', 'unsportsmanlike']
```

Figure 1: Reference Words for Filter

Of these 1,628 comments, it was necessary to understand which comments were posted on the Reddit Game Thread were about a specific penalty called during the game. For this, we created timestamp “buckets” to group each comment within the time that passed between penalty calls. 13 of these buckets were created to sort them into comments that were made before the game started, and for each of the 12 penalties that were called throughout the game. The UNIX timestamps were already collected for each comment while importing from Reddit. The bucket timestamps were also already collected from the play-by-play tweets and conditions were then given to determine which comments were placed in which bucket (see Figure 3). Since the tweets were in real-time, a real-time to UNIX time converter was created within the script to match the tweet timestamps to the comment timestamps since they were not initially compatible to run a comparison. Then, each bucket had its own specific UNIX time, and comments that were posted after the bucket timestamp were placed in the bucket. This would be repeated for all comments until the next bucket timestamp appeared, where comments would be placed there instead. Any comment that occurred before the first penalty was placed in a bucket as well.



```
def real_to_unix(realtime):
    year = 2021
    month = 1
    day = 17
    partitioned_realtime = realtime.partition(':')
    hour = int(partitioned_realtime[0])
    minute = int(partitioned_realtime[-1])
    datetime_real = datetime(year, month, day, hour, minute)
    unixtime = time.mktime(datetime_real.timetuple())
    return unixtime
```



```
# assign comment to penalty bucket
comments_penalty_bucket = np.zeros_like(comment_date_timestamps)
i = 0
while i < len(comment_date_timestamps):
    j = 0
    while j < len(penalty_unixtime_uniq):
        if comment_date_timestamps[i] > penalty_unixtime_uniq[j]:
            comments_penalty_bucket[i] = j
        j += 1
    i += 1
```

Figure 2: Converting to UNIX time

Figure 3: Assigning Comments to Buckets

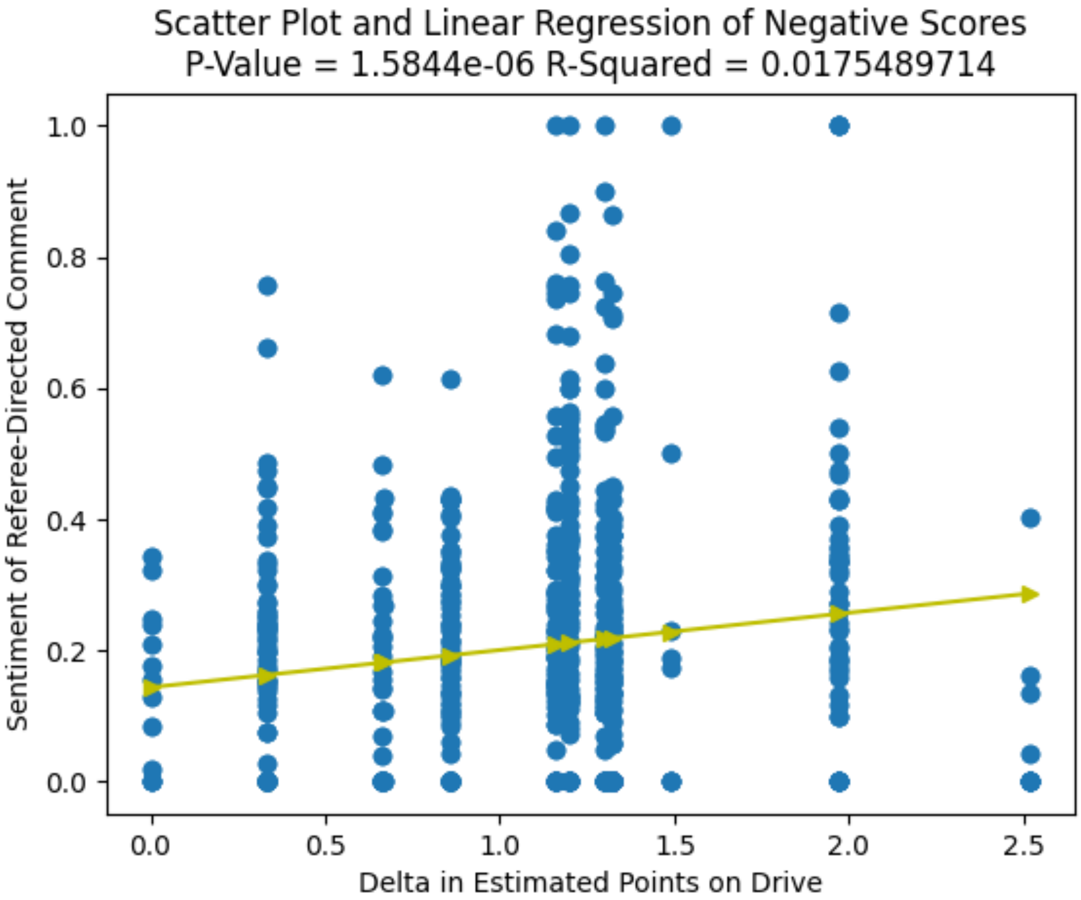
Once the comments are in their respective buckets, VADER sentiment analysis (Hutto & Gilbert, 2014) will be performed on the body of each comment using the Python script. VADER sentiment analysis gives a score between -1 and +1 for the positivity, neutrality, and negativity of each comment. VADER sentiment analysis also offers a compound score that compares both positive and negative scores in the comment body and gives an overall score to the comment. Scores that are closer to -1 are extremely negative and scores that are closer to +1 are extremely positive. Given that these comments are in their respective buckets, the scores on each comment are a direct result of the penalty that was called on that play. The resulting score of each comment will be added alongside the body, timestamp, and team flair keys in the JSON comment structure. The team flairs were initially imported to tie each fanbase to an expected sentiment change but were not used to generate conclusions for this study because it was not needed. For the research that was conducted, the positive, neutral, negative, and compound scores were stored for analysis.

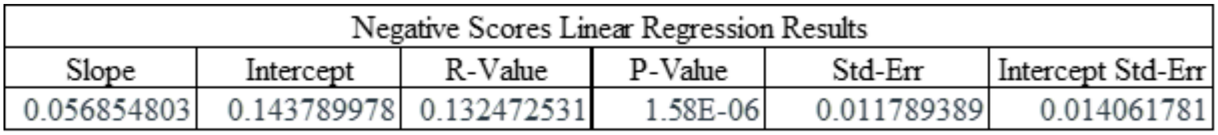
With the data collected and properly structured, the sentiment of comments will be plotted against the expected points delta of the corresponding penalty call. Using these independent and dependent variables of sentiment scores and the expected points delta, linear regression will be performed to determine the correlation of the two variables. The results will give the ability to conclude the expected language and dangers associated with the language that referees experience when calling big plays in football. If the data shows that the two variables are correlated, the conclusions drawn from the analyzed data will be extremely accurate and held to a high standard. The linear regression graphs are shown in the results section below.

Results

Using the Reddit Game Thread from the NFC Divisional Game between the Tampa Bay Buccaneers and the New Orleans Saints, there were over 45,000 comments collected from the Reddit API Scraper. Those comments were then passed through the curated filter of keywords that were relevant to referees calling penalties during the game which returned 1,628 comments. The 1,628 comments were then placed into the penalty timestamp buckets in which every comment had its own assigned bucket based on the penalty that was called. VADER sentiment analysis is run on each comment and given a score that was assigned to each comment. Along with each comment's stored sentiment score was the called penalties expected points delta which was also stored with the comment. Both scores were used as independent and dependent variables which resulted in positive, neutral, negative, and compound linear regression graphs.

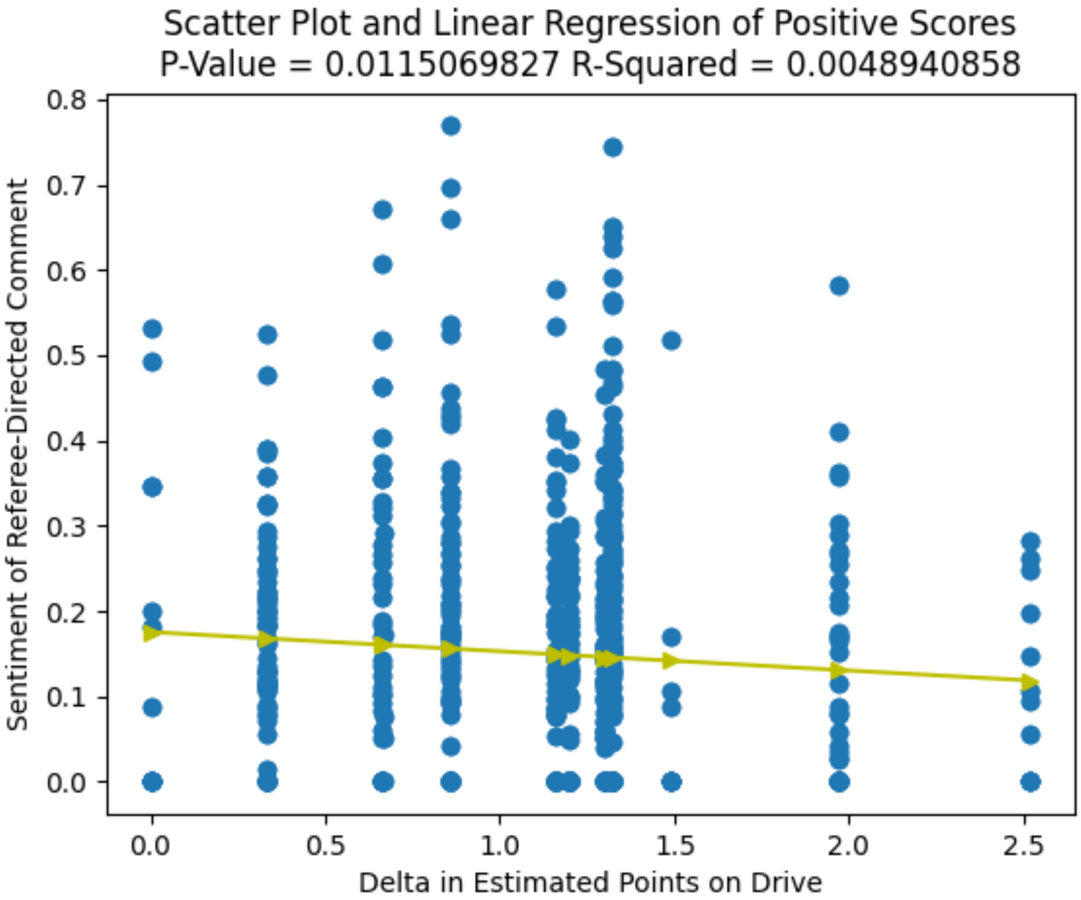
This study calls for the observation into the negative and positive reactionary language directed towards referees. Consequently, the compound and neutral sentiment could be discarded as results are obfuscated by a negative and a positive combining into a neutral. The positive and negative graphs are shown below:

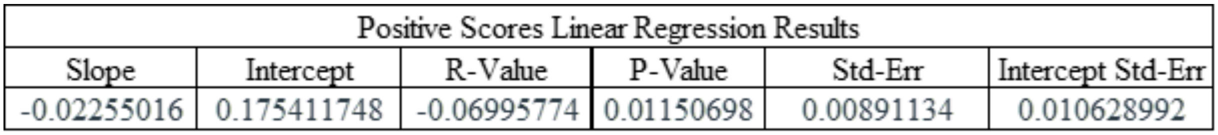




Negative Scores Linear Regression Results					
Slope	Intercept	R-Value	P-Value	Std-Err	Intercept Std-Err
0.056854803	0.143789978	0.132472531	1.58E-06	0.011789389	0.014061781

The graph above shows the linear regression for negative scores throughout the duration of the Divisional Game. The table shows the numeric results of the negative sentiment scores based on the linear regression graph. The slope of 0.0568 shows that as the expected points delta increases, the negative sentiment towards the penalty called increases as well. Other values shown that are important to note are the R^2 - value and the P-Value. The P-Value is extremely low, 0.00000158, which is well below 0.05, meaning that we reject the null hypothesis and conclude that the data is well correlated with one another. R^2 shows the percentage of variation in negative sentiment scores as explained by the Delta in Estimated Points on the Drive. In this instance, the R^2 - value is approximately 0.0175 which indicates that while the two variables are strongly correlated, the linear regression does not create a reliable predictive model. The standard error of approximately 0.0118 indicates the reliability of the mean. A small standard error indicates that the sample mean shown in our data is a more accurate reflection of the total population mean. The linear regression results show that the two parameters are very correlated and therefore reasonable conclusions can be drawn from the data set.





Positive Scores Linear Regression Results					
Slope	Intercept	R-Value	P-Value	Std-Err	Intercept Std-Err
-0.02255016	0.175411748	-0.06995774	0.01150698	0.00891134	0.010628992

The graph shown above represents the linear regression of positive scores for the NFC Divisional Game. The table shows the numeric results of the linear regression positive scores. The slope of -0.0226 shows that as the delta in estimated points on drive increases, the positive sentiment in each comment towards referees decreases. Much like the negative scores, the P-Value of 0.0115 is low as well. While it is not as low as the negative scores, the P-Value for positive scores is below 0.05 as well, rejecting the null hypothesis and concluding that the data is well correlated. The R^2-value of 0.0049 shows the percentage of variation in negative sentiment scores as explained by the Delta in Estimated Points on the Drive is again extremely low. This again indicates that while the two variables are strongly correlated, the linear regression does not create a reliable predictive model. The standard error of 0.0089 shows how reliable the mean of the data is. Since the standard error is very small, the sample mean shown in our data is a more accurate reflection of the total population mean. The linear regression results for the positive results are strongly correlated and verify that conclusions can be drawn from the data.

The results of the data show that both linear regression graphs resulted from sound data collection methods and the parameters dependent and independent variables used were strongly correlated. Given the results from the graphs, reasonable conclusions about the research question can be made from the curated data. When there is a significant penalty call that affects the game, there will be more reactionary language both positive and negative directed towards referees. While there is more reactionary language for both positive and negative, the data shows that negative language is doubled that of positive language. This is seen by comparing the slope for each graph. The negative sentiment slope is 2.5 times greater in magnitude compared to the positive slope. This indicates that when an important call is made on the field, the reactions will most likely be negative towards said call.

Discussion

The results of the data from the linear regression graphs above strongly suggest that the data we chose to compare are not only accurately compiled, but strongly correlated as well. For the 1,628 comments that were used in this study, we can draw reasonable conclusions due to how convincing the results from the linear regression graph were. Not only will this data allow us to answer the research question, but we can also make predictions as to how fans will react when a penalty is called and potentially give insights into keeping referees safer as the game of football continues to grow.

The research question that drove this study was, "How strongly does a play's change in expected points probability indicate the sentiment of messages directed at referees?" Given the results of the data, we can conclude that a play's change in estimated points delta indicates that the negative sentiment towards referees increases while positive sentiment towards referees decreases. This is shown through the linear regression graphs above as well as the slopes that are shown in the tables below each graph. The negative sentiment graph shows a slope of 0.0568, which portrays an increase in negativity towards referees as the estimated points delta increases on penalties called during a game. Contrary to the negative sentiment scores, the positive graph and table reveal a slope of -0.0226, indicating that as the estimated points delta increases, the positive sentiment towards referees decreases. While this could logically be deduced from attending raucous sports events, the analysis now confirms the theory.

Big plays that have a larger estimated points delta will draw more outrage from a fan base. One example of this is in the NFC Divisional Game used for this study where a penalty negated a touchdown from being scored. Before the penalty was called, the Saints Special Teams returned a 48-yard punt for a touchdown which is one of the rarer forms of scoring in Football. However, there was a penalty called on the Saints for an Illegal Block in the Back which negated the touchdown causing fans to turn irate both at the referee and the team. This play had a large estimate points delta and received some of the largest negative sentiment comments on Reddit towards the referees of the entire game. Our study supports this example and answers the research question by showing that the change in estimated points delta is a very strong indicator of both positive and negative sentiment towards referees.

Punt return touchdowns nullified by penalty are not tracked in play-by-play data. Therefore, when collecting estimated points delta on the aforementioned play, it is noted as a minor swing rather than the correct much larger swing. This is the only penalized play in the game that was affected by this data shortcoming. If the true change in estimated points from the call is considered, the estimated point delta of the penalty call is 6.04 rather than 1.2. Using the prediction model generated from the linear regression, we can predict that the population mean of negative referee sentiment would be approximately 0.4689 as calculated by VADER.

It is important to note that due to the low standard error, the prediction model can reliably predict mean population sentiment. However, the low R^2-value indicates that the model cannot reliably predict any single person's sentiment. This is visually shown in the linear regression graphs that show that the regression line correlates between the values, but the data points are not a close fit to the line. There are stronger factors at play when determining the reactive negative sentiment towards a penalty, such as an individual person's temperament.

One interesting analysis that was made from the data above is the difference in change between the positive and negative sentiment as estimated points delta increases. This can be shown by taking the magnitude of the slope. For the positive sentiment, the magnitude of the slope is 0.0226 while the magnitude value for the slope of negative sentiment is 0.0568. Given the results of the negative sentiment slope, we can see that the negative sentiment is 2.5 times greater than the positive sentiment towards referees. This is an interesting discovery because the intensity of the reactions is drastically different between positive and negative sentiment. The resulting difference in slopes helps to show that the larger the penalty affects the estimated points in the game, the outrage towards the referees will tend to be 2.5 times that of the support.

Due to the perceived negative attention from wronging a team via a penalty, refs may try to minimize their total effect on estimated points to minimize negative sentiment. "These calls are intended to atone for a previous error" and "no referee wants to be in the position of having decided a hard-fought contest by denying a team a touchdown or by giving possession of the ball to the wrong team as the result of a refereeing error" (Pizzi, 2008). However, due to plays creating more negative than positive, the "make up calls" create greater total negative sentiment and as a result, more danger. Regardless of if the referee succeeded in neutralizing their effect on the estimated points, their increased activity will only draw further outrage.

This tends to be the case across many sports around the world, not just the NFL. Whenever there are penalties called during a sporting event, there will always be outrage towards the referees for either slowing down the game or making what they believe to be the wrong call. This is especially evident when the estimated points delta is involved as well. When fans notice that the points delta has drastically changed more than other plays due to a penalty called by a referee, they tend to be more outraged than those who aren't. These outraged fans will make their opinions of the calls and the referees very prominent and a useful platform for people to share their outrage amongst themselves is Reddit. People will tend to look for those who are fans of their own team to confirm their outrage and share it with other upset fans who share the same disapproval of the change in the estimated points delta. These instances tend to draw negativity and support it on platforms like Reddit.

However, unlike the outraged fans who disagree with the situation, fans who have positive sentiment towards the change in estimated points delta are not as vocal about their approval of the change. These fans are not using Reddit to share their approval of the change nearly as much as those who are upset and that is shown by the difference in slopes. These fans might not be sharing their approval for many reasons but the platform itself plays a large role in the reaction of the fans. Fans who approve of the change in estimated points delta may not want to voice their opinion because of the backlash they may feel from the outraged fans. People who are outraged tend to not rationalize opinions besides their own which may deter those who have positive sentiment to the call.

Another reason why the fans may not be using Reddit to voice their approval is the positivity felt towards the play may not lead them to voice their opinions on the platform. Given that these fans already feel happy about the call, they may not express their feelings because there is no need to feel better about it by looking for other fans who are just as happy as they are. This will cause the number of comments that are positive towards big plays to decrease such as the data suggests. This helps add context as to the reasoning behind negative comments doubling as the change in estimated points delta increases compared to the positive comments.

It is important to note that while the study itself offers many conclusive results and allows for reliable interpretations of the data to be made, several potential errors that occur with the VADER Sentiment Analysis, as well as the data collection, may impact the data. Most notably, VADER analysis does not recognize sarcasm in text-based comments. VADER analysis has a very difficult time detecting any sarcasm because of its lack of contextual evidence as well as the knowledge of what the sarcasm is directed towards. This can cause some of the data points from the comments to not be as accurate as others that are scored correctly.

Another flaw with the VADER sentiment analysis is negative language receiving incorrect scores due to a misinterpretation of the modern slang that is used in each comment. Derogatory plays on words that are directly targeting the referees may not be counted as negative comments and will be given either neutral or positive scores. These comments are in the wrong group and could potentially skew the results.

In regard to data collection, one flaw that was noticed in the collection of comments was the inability to import all the comments at once. Neither one of us had access to Reddit premium which offers the ability to open every comment at once. This meant that to retrieve every comment, the requirement to import comments by old and again by new was necessary to retrieve every comment within the subreddit. This allowed for a negligible, but existent, number of duplicate comments to be in the data set for analysis. While our script addressed these potential errors, the chance for a few duplicate comments skewing data still exists and should be noted. However, we determined that even with VADER sentiment not having the greatest accuracy and the possibility of a few comments having errors in the import process, the 1,628 data points collected were enough to negate any potential errors and therefore allowing for us to draw accurate conclusions based on the data.

Conclusion

From importing the over 45,000 comments of the NFC Divisional Game versus the Tampa Bay Buccaneers and the New Orleans Saints and filtering them down to 1,628 relevant comments, accurate conclusions about the data can be made to answer the research question posed. This study aimed to find out whether a play's change in expected points probability is an indicator of the sentiment of messages directed at referees. Based on the data, it is an accurate assumption to say that the play's change in expected points is an indicator of the sentiment of the messages that will be directed towards the referees. If the play causes the drive’s estimated points to increase, then the negative sentiment will be doubled that of the positive sentiment towards referees for that play. This data allows for opportunities for continued research into the future that may help increase referee safety based on fan reactions to their penalty calls, similar to the proposed theory of makeup calls only further risking referee safety. Referees can benefit greatly from this information and the probability of seeing violent incidents will vastly decrease allowing referees to feel much safer making penalty calls.

References

Abc7LosAngeles, director. *Texas High School Football Player Attacks, Knocks over Referee after Being Ejected from Game | ABC7. YouTube*, YouTube, 4 Dec. 2020, www.youtube.com/watch?v=b6s4CwjjMFA&ab_channel=ABC7.

Boe, Bryce. "The Python Reddit API Wrapper." *PRAW*, 2021, praw.readthedocs.io/en/latest/.

"Divisional Round - Tampa Bay Buccaneers at New Orleans Saints - January 17th, 2021." *Pro*, www.pro-football-reference.com/boxscores/202101170nor.htm#all_pbp.

Hutto, CJ, and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

Jacobs, Brittany L., et al. "Exploring Referee Abuse through the Lens of the Collegiate Rugby Coach." *Sport Management Review*, vol. 23, no. 1, 2020, pp. 39–51., doi:10.1016/j.smr.2019.03.004.

Pizzi, William T., 'Makeup Calls' in Sports and Courts (June 22, 2008). The Green Bag 2D, Vol. 11, p. 333, Spring 2008, Available at SSRN: https://ssrn.com/abstract=1331635

Reddit. "r/Nfl - Game Thread: Tampa Bay Buccaneers (11-5) at New Orleans Saints (12-4)." *Reddit*, 2021, www.reddit.com/r/nfl/comments/kzhfxm/game_thread_tampa_bay_buccaneers_115_at_new/.

Scoppa, Vincenzo. "Social Pressure in the Stadiums: Do Agents Change Behavior without Crowd Support?" *Journal of Economic Psychology*, vol. 82, 2021, p. 102344., doi:10.1016/j.joep.2020.102344.

Twitter, Twitter, twitter.com/bucs_nation?lang=en.

"What Are The Most Common Penalties In NFL Football?" *Rookie Road*, Rookie Road, 11 Apr. 2021, www.rookieroad.com/football/questions/what-are-the-most-common-penalties-in-nfl-football/.